Discriminative Multi-instance Multitask Learning for 3D Action Recognition

Yanhua Yang, *Member, IEEE*, Cheng Deng, *Member, IEEE*, Shangqian Gao, Wei Liu, Dapeng Tao, *Member, IEEE*, and Xinbo Gao, *Senior Member, IEEE*

***Abstract*—As the prosperity of low-cost and easy-operating depth cameras, skeleton-based human action recognition has been extensively studied recently. However, most of the existing methods partially consider that all 3D joints of a human skeleton are identical. Actually, these 3D joints exhibit diverse responses to different action classes, and some joint configurations are more discriminative to distinguish a certain action. In this paper, we propose a discriminative multi-instance multitask learning (MIMTL) framework to discover the intrinsic relationship between joint configurations and action classes. First, a set of discriminative and informative joint configurations for the corresponding action class is captured in multi-instance learning model by regarding the action and the joint configurations as a bag and its instances, respectively. Then, a multitask learning model with group structure constraints is exploited to further reveal the intrinsic relationship between the joint configurations and different action classes. We conduct extensive evaluations of MIMTL using three benchmark 3D action recognition datasets. Experimental results show that our proposed MIMTL framework performs favorably compared with several state-of-the-art approaches.**

***Index Terms*—3D action recognition, discriminative multi-instance learning, group sparsity, joint configuration, multitask learning.**

I. Introduction

Human action recognition is a major component of many computer vision applications, e.g., intelligent surveillance, human-computer interfaces, health care, and virtual reality. Despite considerable progress has been made in the last decade [1]–[8], recognizing human actions accurately in video sequences remains difficult. It is mainly because that occlusions, background clutter, illumination changes, individual style, and viewpoint changes can disable the entire capturing of human actions, leading to large intra-class variations within action classes and inter-class ambiguities between action classes. To address the above mentioned issues, recently a simple yet effective way to recognize human actions, is introduced with the advent of the depth sensor. Compared to monocular camera, depth sensor can provide real-time estimation of 3D joint positions of a human skeleton, thus the influence of cluttered background and illumination variations can be eliminated [9]. Meanwhile, Shotton *et al.* [10] proposed a rather powerful human motion capturing technique that estimates 3D joint positions of a human skeleton from a single depth image. Therefore, ever-increasing human action recognition methods are reported to focus on the sequences of 3D joint positions.

Given the corresponding 3D joint positions, an action can be intuitively represented as a temporal sequence of approximated human skeleton. Existing joint-based action recognition approaches are to model the dynamics of either single joint or combination of joints using various features, e.g., joint positions (JP) [11], joint orientations (JO) [12], and pairwise relative joint positions (RJP) [13]. However, it is observed that not all of human joints can effectively reflect the dynamic variation of the corresponding action. That is to say, some joints are more relevant to the action, while others not. Although being the action element, a single joint is unqualified to recognize the corresponding action, because it has no discriminant ability of distinguishing different action classes. Many literatures [14]–[16] have proven that one of the important factors in visual recognition is to filter the training data. The noisy training data, such as inaccurate data labels and useless data, may degrade the performance of the learning model. Therefore, it is necessary and important to discover a set of informative joint configurations (i.e., possible joint subset consisting of two or more joints) for human action recognition.

However, automatic discovering a set of discriminative joint configurations is extremely difficult, mainly due to two problems. First, the joint configurations cannot be manually defined because the number of joint configurations is enormous for all action classes. Second, it is hard to establish the intrinsic relationship between the joint configurations and a specific action class, where the discriminant ability should be accurately defined to characterize the response degree of the joint configurations to the corresponding action. Considering the fact that an action contains a set of different joint configurations, the

Manuscript received May 1, 2016; revised July 26, 2016 and October 23, 2016; accepted October 23, 2016. Date of publication November 9, 2016; date of current version February 14, 2017. This work was supported by the National Natural Science Foundation of China under Grant 61572388 and Grant 61432014, by the National High Technology Research and Development Program of China under Grant 2013AA01A602, by the Key Industrial Innovation Chain Project in Industrial Domain under Grant 2016KTZDGY-02, by the Program for Changjiang Scholars and Innovative Research Team in University under Grant IRT13088, and by the Shenzhen Technology Project JCYJ20140901003939001. The guest editor coordinating the review of this manuscript and approving it for publication was Prof. Yingcai Wu.

Y. Yang, C. Deng, W. Liu, and X. Gao are with the School of Electronic Engineering, Xidian University, Xi’an 710071, China (e-mail: yangyanhua.xd@gmail.com; chdeng@mail.xidian.edu.cn; weiliu.xidian@gmail.com; xbgao@mail.xidian.edu.cn).

S. Gao is with the College of Engineering, Northeastern University, Boston, MA 02115 USA (e-mail: gao.sh@husky.neu.edu).

D. Tao is with the School of Information Science and Engineering, Yunnan University, Kunming 650091, China (e-mail: dapeng.tao@gmail.com).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TMM.2016.2626959

joint configurations and the corresponding action class can be viewed as a bag and its instances, respectively. Thus, we can discover the joint configurations with multi-instance learning (MIL), in which the labels are assigned to the instances (i.e., joint configurations) when the labeled bags (i.e., action classes) were known.

In real-world scenarios, two actions may be only differentiated by very subtle human skeletal details, which requires the captured joint configurations having more discriminant abilities. Moreover, since many action classes are highly correlated, leveraging the intrinsic correlation across different actions can significantly enhance the recognition performance. Accordingly, a set of good joint configurations should be able to generalize over variations within one class as well as distinguish between different action classes. Inspired by multitask learning (MTL), we consider each action as a learning task and exploit the interdependencies across different actions to further enhance the discriminant abilities of the joint configurations.

In this paper, we propose a two-stage learning framework, termed multi-instance multitask learning model (MIMTL), for 3D human action recognition. Specifically, in the stage of MIL, each action sequence is considered as a bag of instances, in which each instance corresponds to a possible joint configuration. Then MIL paradigm is applied to build the relevant relationships between the bags (i.e., actions) and the instances (i.e., joint configurations), where the labels of joint configurations can be completely inferred when the action classes information are known. According to the consistencies between the labels of joint configurations and the action classes, the most discriminative joint configurations for a specific action class can be determined. In the stage of MTL, each action is viewed as a task, and all actions are recognized simultaneously by exploiting the knowledge shared among actions. Traditional assumption in MTL model is that all tasks should share some common structure, which is too restricted in real-world applications. Since the tasks have the property of implicit group structure, we utilize complex sparsity constraints to encourage the actions to reveal this property. Thus, the discovered joint configurations have better discriminant ability to resist the intra-class variations within action classes and the inter-class ambiguities between action classes.

In summary, the major contributions of this paper include: 1) This paper proposes a multi-instance multitask learning model as a flexible way to characterize and recognize 3D human actions; 2) we exploit multiple instance learning paradigm to discover the most discriminative joint configurations for a specific action class by considering one action sequence as a bag of instances; 3) we infer the intrinsic correlations among different actions based on the discovered joint configurations with a task-grouping regularization in multitask learning paradigm.

Our proposed method is evaluated on three benchmark datasets, namely MSR-Action3D [17], UTKinect-Action dataset [12] and Florence3D-Action dataset [18]. Extensive experiments show the advantages of our method on 3D action recognition comparing with other state-of-the-art approaches.

The remaining sections are organized as follows. Section II reviews related works. In Section III, we provide the formulation of the proposed method and the corresponding optimization algorithm for our MIMTL model. Comprehensive experimental results on several datasets are presented to demonstrate the effectiveness of the proposed method in Section IV. Finally we conclude in Section V.

II. Related Work

In this section, we briefly review the works related to our proposed approach which involves three important components: skeleton-based action recognition, multi-instance learning, and multitask learning.

Skeleton-based action recognition: In 3D skeleton-based action recognition, an action is described as a collection of time series of 3D positions of the joints in the human skeleton. In general, existing works can be divided into two categories: joint based action recognition and mined-joint based action recognition.

Most of existing works belong to the first category, in which various features are extracted from the skeleton data in order to capture the correlation of the body joints existed in spatial or geometric space. In [19], an action was modeled as a spatio-temporal structure, in which a set of 13 joint positions in a 4-D space (3D+time) was exploited to describe a human action. In [20], Lv and Nevatia designed seven types of feature vectors, such as joint position, bone position, joint angle, and so on, and then learned the dynamics of each action via HMM for each feature, respectively. Chaudhry *et al.* [21] proposed a spatio-temporal hierarchy of skeletal configuration, where each configuration represents the motion of a set of joints at a particular temporal scale. In [11], the covariance matrix for joint positions in a fixed length temporal window was utilized to capture the relations among joints. In [22] a generic kernel matrix feature representation was proposed to elevate the limitation of covariance representation. Wang *et al.* [13] captured all the pairwise relative positions of the joints to characterize the spatial relationship of joints. However, these methods ignored the temporal information completely, leading to ambiguity description of action sequence. In [23], 3D position differences of the same joints between different frames were extracted to characterize the dynamics. The work in [24] proposed a more complex skeleton representation. By describing the relative geometry between two rigid body parts as a rigid rotation and translation transformation, a sequence of skeleton was represented as a curve in the Lie group. In [25], each action was represented by spatio-temporal motion trajectories of the joints. Trajectories are curves in the Riemannian manifold of open curve shape space. Sharaf *et al.* [26] introduced a real-time multi-scale action detection system for 3D skeleton data. In [27], a moving pose descriptor that considers both pose information and differential quantities of the human body joints were computed. However it required providing a label for every frame during training and testing which was not practical. In general, the methods in this category usually design elaborate hand-crafted features followed by classifier without joint selection, and all joints engaged in the action are considered to be made the same contribution to the action recognition.

In contrast to joint based action recognition, mined-joint based action recognition assumes that detection of the activated subsets of joints is beneficial to distinguish different action classes. Methods in this category focus on mining the subsets of the most discriminative joints. In [6], the most informative joints were discovered by calculating the relative informativeness of all the joint angles based on their entropy, where the joint that has the highest variance of angle change was defined as the most informative one. In [28], the informativeness of each joint for certain action class was determined based on the degree of variation in its position using the information entropy. And a combination of spatio-temporal based skeleton features was employed and longest common subsequence (LCSS) algorithm was utilized as similar function to determine the action class. Wei *et al.* [29] adopted multiple kernel learning to mine the most informative joints, and then trajectories of difference vectors between the selected joints were utilized as the action representation. The approach in [30] employed partial least squares to weight the importance of each joint. In [31], the importance of each joint was measured from the information theory point of view, in which the more discriminative joint subsets were selected by calculating the mutual information (MI) between each joint and an action.

Multi-instance learning: Many techniques involving multi-instance learning (MIL) can be found in the literature and we refer readers to the review [32]. MIL was first introduced in the context of drug design by Dietterich *et al.* [33]. Besides the drug activity prediction, MIL had been applied to other challenging computer vision learning problems, including image indexing for content-based image retrieval [34], scene classification [35], and video event detection [36]. The two most popular algorithms of MIL are mi-SVM and MI-SVM. mi-SVM emphasizes searching max-margin hyperplanes to separate positive and negative instances, while MI-SVM aims to select the most representative positive instance for each positive bag during optimization iterations, and then concentrates in bag classification.

The traditional MIL was described in a scenario like: each training data consists of a bag of instances. The labels are only provided on the bag-level and the instance labels are unknown. It is usually assumed that a positive bag contains at least one positive instance, while all of the instances in a negative bag must be negative. This is a very weak assumption and has been proved not to be suitable to the real-world applications. A number of impressive approaches have developed with different assumptions [36]–[38]. In [36], MIL was applied to video event detection by representing each video as multi-granular temporal video segments, and the key assumption was that a large portion of instances in a positive video should be positive, whereas few instances in the negative videos may be positive. In [37], Li *et al.* formulated the image retrieval task as a MIL problem by assuming that each positive bag contains at least a portion of positive instances. The work of [38] encoded cardinality relations directly by scoring a clip with the number of interesting frames it contains and a novel kernel was proposed for modeling cardinality relations, counting instance labels in a bag.

Multitask learning: multitask learning (MTL) aims to jointly learn several tasks by leveraging the correlations among multiple related tasks. There are various assumptions. One common assumption is that all the tasks are related [39]–[41], which does not work in practice. Negative transfer would be inevitable if unrelated tasks exist. Thus more complex approaches have been proposed to address this problem. Some works assumed a *priori* knowledge, such as tree-guided MTL [42], clustered MTL [43], [44], and GO-MTL [45]. However, most *priori* knowledge are pre-defined [46], which may be unreasonable in realistic setting. For action recognition scenarios, Mahasseni and Todorovic [47] learned latent action groups to achieve view-invariance in action recognition, while within-group feature sharing was allowed but between-group feature sharing was prohibited. In [48], the latent tasks corresponding to basic motion patterns were learned across action classes. However, the so-called basic motion pattern is implicit, so that the learned latent tasks have no physical meaning. Lin *et al.* [49] augmented target action representation with additional depth and skeleton information by using domain adaptation learning. The method proposed in [50] tried to train one-vs-one SVM classifiers for every pair of categories to compute the category-level similarity. The work of [51] constructed the super-category by measuring the similarity of action categories based on mutual information.

Similar to mined-joint based action recognition, our purpose is also to mine a set of discriminant joints. Differently, our proposed scheme learns a set of possible joint configurations with MIL model instead of utilizing the hard-decision information entropy in previous approaches. Furthermore, although existing MTL-based action recognition approaches build the relationship between visual features and action classes, our proposed two-stage method discovers the intrinsic relationships not only between joints configurations and action classes but also among action classes with complicated task-grouping regularization.

III. The Proposed Approach

In this section, we will introduce our discriminative MIMTL framework for 3D action recognition in detail. In general the proposed method consists of two learning stages. The first stage aims to discover a set of discriminative joint configurations in the multi-instance learning paradigm. And in the second stage, multitask learning paradigm is employed to boost recognition accuracy by discovering the intrinsic relationship between the joint configurations and the different action classes. Finally, we present the optimization algorithm for our proposed framework.

A. Notation Definition

Suppose we are given a training set of input-output structure pairs $\mathcal{T} = (\mathbf{X}, \mathbf{Y}) = \{(X_{il}, Y_{il})_{i=1}^{n_l}\}_{l=1}^{L}$, where each X_{il} represents a single action sequence, $Y_{il} \in \{1, 2, \ldots, L\}$ is the corresponding action class, L is the number of action classes, n_l is the number of samples in l-th class, and $N = \sum_{l=1}^{L} n_l$ is the total number of the training data. The pairwise relative joint position (RJP) is used as the feature representation. Specifically, we employ the 3D coordinates difference between some joint and other joints to characterize the action information. For each action X_{il}, the 3D joint positions are arranged in time sequence.

B. Discriminative Multi-instance Learning

In the original MIL model, each training sample is viewed as a bag where all unlabeled instances in this bag share a common label. For example, an image is usually considered as a bag of multiple local patches, and a video can be regarded as a bag of multiple frames. For 3D action recognition, in analogy with MIL, we consider each single action sequence X_{il} as a labeled bag, and some possible joint configurations related to this action as the instances. Different from original MIL model aiming to predict the labels of unseen bags, however, our MIL model is only utilized to infer the labels of the instances of the labeled bag.

Since the human skeleton is always represented by a hierarchy of joints, we define each joint configuration as a combination of human joints. In the pioneer work [31], such a joint configuration is dubbed as *skelets*. In our proposed approach, each action sequence is represented by a set of discriminative joint configurations which will be discovered in MIL model.

Suppose there are N_J joint configurations, we denote $X_{il} = \{x_{il}^m, y_{il}^m\}_{m=1}^{N_J}$, in which x_{il}^m means the m-th instance of the i-th sample belonging to the l-th bag, $y_{il}^m \in \{-1, 1\}$ is the corresponding instance label. $y_{il}^m = 1$ if the instance is positive for this bag, otherwise $y_{il}^m = -1$. In fact, the instance labels are unknown. The supervised information is only provided on bag-level. Even so, we still can classify instances in a bag according to their contributions to the success classification for a specific action class. For a specific action class, only a few joint configurations are related to it, which means these joint configurations have good discriminant ability to sufficiently represent this action class. Therefore, the discriminant ability of m-th joint configuration for l-th action class is defined as

$$d_l(\overline{\mathbf{y}}^m) = \frac{\overline{\mathbf{Y}} \cdot \overline{\mathbf{y}}^{m\,T}}{N_l}. \quad (1)$$

Here N_l is the number of training data for the l-th action class ($N_l \ll N$). $\overline{\mathbf{Y}} = [\overline{Y}_1 \cdots \overline{Y}_j \cdots \overline{Y}_{N_l}]$ is the corresponding bag labels, in which $\overline{Y}_j = 1$ if the j-th data belongs to the l-th action, otherwise $\overline{Y}_j = -1$. $\overline{\mathbf{y}}^m = [\overline{y}_1^m \cdots \overline{y}_{N_l}^m]$ is the m-th instance label for all N_l data, whose element equals to 1 or -1. As defined in (1), the discriminant ability of joint configuration can be evaluated efficiently with the consistency between the bag labels and the inferred instance labels. The maximum value is 1 when the instance labels are the same with the corresponding bag labels. From the perspective of MIL, such instance must be the positive for the positive bag or negative for the negative bag. Thus the corresponding joint configuration appears to be a discriminative joint configuration for the l-th action class.

Besides the discriminant ability, the informativeness of each joint configuration is also considered. We define the informativeness as the proportion of the positive instance for l-th bag as follows:

$$p_l^+(\overline{\mathbf{y}}) = \frac{1}{N_J} \sum_{m=1}^{N_J} \mathrm{I}(\overline{\mathbf{y}}^m = \mathbf{1}) \quad (2)$$

where $\mathrm{I}(\cdot)$ is the indicator function which is 1 even if only one element in $\overline{\mathbf{y}}^m$ is 1, otherwise 0. The value of $p_l^+(\overline{\mathbf{y}})$ is the range of $(0, 1]$. The smaller the value, the larger the informativeness. If the value is equal to 1, it means all of N_J joint configurations are always positive for this action class. That is to say, each joint configuration contains different informativeness, and those joint configurations with large informativeness have more powerful ability of characterizing the corresponding action.

Subsequently, we exploit MIL model to infer the labels of the instances. The objective function is defined as

$$\min_{y_{il}^m, \mathbf{w}_l, b} \frac{1}{2}\|\mathbf{w}_l\|^2 + C \sum_{i=1}^{N_l} \sum_{m=1}^{N_J} \mathcal{L}(y_{il}^m, (\mathbf{w}_l^T x_{il}^m + b)) + C_d \sum_{m=1}^{N_J} |1 - d_l(\overline{\mathbf{y}}^m)| + C_p p_l^+(\overline{\mathbf{y}}). \quad (3)$$

The first term in (3) is the model parameter, and the second term $\mathcal{L}(\cdot)$ is the empirical loss between the initial instance label and the inferred label. Here, we use the hinge loss function

$$\mathcal{L}(y_{il}^m, \mathbf{w}_l^T x_{il}^m + b) = \max(0, 1 - y_{il}^m(\mathbf{w}_l^T x_{il}^m + b)). \quad (4)$$

The third term in (3) is used to encourage the discriminant ability of each joint configuration, with which the most discriminative joint configurations are explored. And the last term constrains the number of the joint configurations. Not all joint configurations are expected to be selected as the discriminative ones. C_d and C_p are the parameters to balance the constraints.

In practice, we apply the alternating optimization to find a local suboptimal solution for (3). We first fix instance labels, and then solve $\mathbf{w}_l$ and b. By fixing y_{il}^m, the optimization problem becomes a classic SVM model

$$\min_{\mathbf{w}_l, b} \frac{1}{2}\|\mathbf{w}_l\|^2 + C \sum_{i=1}^{N_l} \sum_{m=1}^{N_J} \mathcal{L}(y_{il}^m, (\mathbf{w}_l^T x_{il}^m + b)). \quad (5)$$

Then, we fix $\mathbf{w}_l$ and b, and update instance labels y_{il}^m. The objective function becomes

$$\min_{y_{il}^m} C \sum_{i=1}^{N_l} \sum_{m=1}^{N_J} \mathcal{L}(y_{il}^m, (\mathbf{w}_l^T x_{il}^m + b)) + C_d \sum_{m=1}^{N_J} |1 - d_l(\overline{\mathbf{y}}^m)| + C_p p_l^+(\overline{\mathbf{y}}). \quad (6)$$

Due to the fact that each instance contributes to the objective function independently, we can optimize an instance at a time in (6). Specifically, we first set all instance labels y_{il}^m to -1, and calculate each empirical loss difference δ_{il}^m by changing each y_{il}^m from 1 to -1. The value of δ_{il}^m is computed as follows:

$$\delta_{il}^m = (1 - (\mathbf{w}_l^T x_{il}^m + b))_+ - (1 + (\mathbf{w}_l^T x_{il}^m + b))_+ \quad (7)$$

where $(\cdot)_+ = \max(\cdot, 0)$. Once all δ_{il}^m are obtained, we determine the instance with maximum difference, and then set its label to 1, and so on. The algorithm stops when the reduction of the objective function is less than a predefined threshold.

C. Joint Configurations Selection

We would like to note that the purpose of MIL model in our approach is not to label unseen bags. On the contrary, it is exploited to discover a set of discriminative joint configurations for a certain action class.

Actually, two types of the joint configurations can be selected by the MIL model. The first type is the discriminative ones that require the instance labels consistent with the corresponding bag labels. The top T discriminative joint configurations are expressed as

$$D_l = \left\{ m | \underset{m \in \{1,\cdots,N_J\}}{\arg\max} \sum_{i=1}^{N_l} \mathbf{I}\left(\hat{y}_{il}^m = \overline{Y}_{il}\right) \right\} \tag{8}$$

where $\hat{y}_{il}^m$ is the inferred instance label.

The second type is the confusing joint configurations that need to be considered due to two factors. On the one hand, overfitting may occur if only the discriminative joint configurations are selected. On the other hand, some joint configurations may mislead the final action recognition. For example, a certain joint configuration may simultaneously produce high responses to many different action classes, which will degrade the recognition performance. Therefore, confusing joint configurations should also be selected, which are denoted as

$$C_l = \left\{ m | \underset{m \in \{1,\cdots,N_J\}}{\arg\max} \sum_{i=1}^{N_l} \mathbf{I}(\hat{y}_{il}^m = 1) \right\}. \tag{9}$$

Unlike the discriminative joint configurations, the confusing ones should be suppressed in the subsequent recognition process.

Finally a new subset $\mathcal{J}$ is constructed by merging D_l and C_l, which is expressed as $\mathcal{J} = \cup_{\{l=1,\cdots,L\}}(D_l \cup C_l)$. We denote there are N_s joint configurations inside $\mathcal{J}$.

Thus far, our proposed approach is able to automatically discover both the discriminative joint configurations and the confusing ones for each action class. In the following subsection, to eliminate the large intra-class variations within action classes and inter-class ambiguities between action classes, we need to further discover the intrinsic relationship between joint configurations and action classes.

D. Multitask Learning With Group Constraints

We formulate action recognition into MTL model in which each action class is regarded as a single task. First, the training data $\mathbf{X}$ should be reconstructed in accordance with the subset $\mathcal{J}$. Only the feature corresponding to N_s joint configurations in $\mathcal{J}$ are concatenated to form the new feature representation $\widetilde{X}$, and the reconstructed training data is denoted as $\widetilde{\mathbf{X}}$.

Given a linear MTL model $\mathbf{W} = [\mathbf{W}_1,\cdots,\mathbf{W}_l,\cdots,\mathbf{W}_L]$, each task is characterized by a parameter vector $\mathbf{W}_l$. For simplicity, $\mathbf{W}_l$ is initialized with the learned parameter $\mathbf{w}_l$, D_l and C_l.

Given the selected joint configuration subset $\mathcal{J}$, we attempt to capture the inherent correlation between joint configurations and action classes in the MTL model. To effectively reveal the correlation, each joint configuration inside $\mathcal{J}$ is considered as a hidden variable h, the objective function is then defined as

$$\min_{\mathbf{W},h\in\mathcal{J}} \sum_{l=1}^{L}\sum_{i=1}^{n_l} \Delta((Y_{il}, h^*), (\hat{Y}_{il}, \hat{h})) + \Omega(\mathbf{W}) \tag{10}$$

where

$$h^* \triangleq \underset{h\in\mathcal{J}}{\arg\max}[\mathbf{W}_l \cdot \psi(\widetilde{X}_{il}, h)]$$
$$(\hat{Y}_{il}, \hat{h}) \triangleq \underset{(Y,h)\in\mathcal{Y}\times\mathcal{J}}{\arg\max}\,[\mathbf{W} \cdot \psi(\widetilde{X}_{il}, h)]. \tag{11}$$

Here, h^* indicates the best hidden joint configuration with which the linear predictor $\mathbf{W}_l$ can predict the input $\widetilde{X}_{il}$ accurately; $(\hat{Y}_{il}, \hat{h})$ express the predicted action label and hidden joint configuration for a given data $\widetilde{X}_{il}$ without the knowledge of ground truth Y_{il}.

However the first term $\Delta((Y_{il}, h^*), (\hat{Y}_{il}, \hat{h}))$ in (10) cannot be directly obtained since the best hidden joint configuration h^* is unseen. Thus an upper-bound $\Delta(Y_{il}, \hat{Y}_{il}, \hat{h}_{il})$ is utilized here instead of $\Delta((Y_{il}, h_{il}^*), (\hat{Y}_{il}, \hat{h}_{il}))$ [52]

$$\Delta(Y_{il}, \hat{Y}_{il}, \hat{h}) = \max_{(\hat{Y},\hat{h})\in\mathcal{Y}\times\mathcal{J}} \left[\mathbf{W} \cdot \psi(\widetilde{\mathbf{X}}_{il}, \hat{h}) + \Delta_{0,1}(Y_{il}, \hat{Y}_{il})\right]$$
$$- \max_{h\in\mathcal{J}} \mathbf{W}_l \cdot \psi(\widetilde{\mathbf{X}}_{il}, h) \tag{12}$$

where $\Delta_{0,1}(Y_{il}, \hat{Y}_{il})$ is zero-one loss, i.e., being 1 if $Y_{il} \neq \hat{Y}_{il}$, and otherwise 0. Obviously, $\Delta(Y_{il}, \hat{Y}_{il}, \hat{h})$ is the difference of two convex functions. For the sake of notational brevity, (12) is rewritten as

$$\Delta(Y_{il}, \hat{Y}_{il}, \hat{h}_{il}) = \max\{0, \xi_{il}\} \tag{13}$$

where $\xi_{il} = f_i(\mathbf{W}) - g_i(\mathbf{W}, h)$. $f_i(\mathbf{W})$ and $g_i(\mathbf{W}, h)$ are two convex functions.

Considering that the action classes have the property of group structure, structured sparsity is used naturally in the MTL model to encourage features with inter-group competition and features with intra-group sharing. The group structure of the action classes $\{G_k\}_{k=1}^K$ can be easily obtained by hierarchical clustering, where K is the group number. Thus the second term of (10) is formulated as

$$\Omega(\mathbf{W}) = \lambda_1 \parallel \mathbf{W} \parallel_F^2 + \lambda_2 \sum_{k=1}^{K}\sum_{h=1}^{N_s} \parallel \mathbf{W}^{G_k,h} \parallel_2 \tag{14}$$

where the first term is Frobenius norm aiming to avoid overfitting. $\mathbf{W}^{G_k,h}$ is a subset specified by both the indices of group structure G_k and the hidden joint configurations. In detail, the ℓ_{21} norm encourages features with intra-group sharing, and ℓ_1 norm induces features with inter-group competition. λ_1 and λ_2 are scalar parameters to balance the sparsity properties.

As a whole, our MTL model is formulated as follows:

$$\min_{\mathbf{W},h\in\mathcal{J}} \sum_{l=1}^{L}\sum_{i=1}^{n_l} \max\{0, \xi_{il}\} + \lambda_1 \parallel \mathbf{W} \parallel_F^2$$
$$+ \lambda_2 \sum_{k=1}^{K}\sum_{h=1}^{N_s} \parallel \mathbf{W}^{G_k,h} \parallel_2 . \tag{15}$$

Algorithm 1: The learning procedure of our proposed method.

```
Input: Training Set $\mathcal{D} = (\mathbf{X}, \mathbf{Y}) = \{(\mathbf{X}_{il}, Y_{il})_{i=1}^{n_l}\}_{l=1}^{L}$,
       Joint configurations $N_J$
1: Map each data $\mathbf{X}_{il}$ into bag of $N_J$ instances
   $\mathbf{X}_{il} = \{x_{il}^m\}$
2: For each action class $l$ do
   Learn MIL model parameter $\mathbf{w}_l$ and infer each instance
   label $y_{il}^m$ according to (3)
   Select $D_l$ and $C_l$ according to (8) and (9)
   end For
3: Generate subset $\mathcal{J} = \cup_{\{l=1,\cdots,L\}}(D_l \cup C_l)$
4: Construct group structure information $\{G_k\}_{k=1}^{K}$
5: Reconstruct input data $\mathbf{X}$ to $\widetilde{\mathbf{X}}$ based on $\mathcal{J}$
6: Learn MTL model parameter $\mathbf{W}$ according to (13)
Output: MTL model parameter $\mathbf{W}^{opt}$
```

TABLE I
SUMMARIZATION OF THREE BENCHMARK 3D ACTION RECOGNITION DATASETS

Dataset	*# J.*	*# Action*	*# Act.*	*# Seq.*	*# F.*
UTKinect-Action [12]	20	10	10	199	[5,74]
MSR-Action3D [17]	20	20	10	516	[13,76]
Florence3D-Action [18]	15	9	10	215	[8,35]

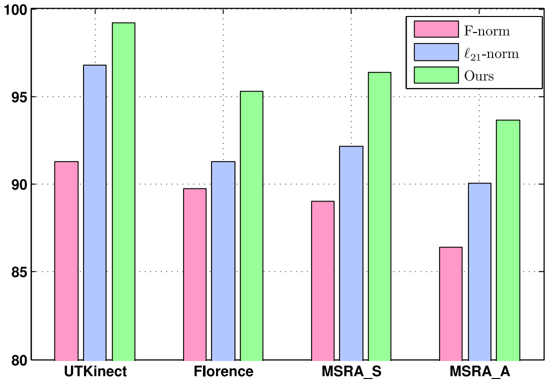


Fig. 1. Effects of different regularization terms on three benchmark datasets.

TABLE II
SUMMARY OF THE STATE-OF-THE-ART APPROACHES USED FOR COMPARISON

Methods	Feature representation	Classifier
Actionlets [13]	RJP & Depth	MKL
HOJ3D [12]	JO	HMM
EigenJoints [55]	RJP	Bayes
Spatial-temporal part-sets [56]	JP	SVM
Cov3DJ [11]	JP	SVM
Random forests [57]	RJP & STIP	Random Forest
Points in a Lie Group [24]	SE3	SVM
LM^3 TL [31]	RJP	MTL
Ker-RP-RBF [22]	JP	SVM
LCSS+MIJA [28]	JO & relative motion	LCSS
Proposed method	RJP	MTL

One thing to note is that the constraints in (15) is a mixed norm regularization which is convex but non-smooth and nontrivial to optimize. Following the steps in [46], [53], the mixed norm is replaced by their dual form, respectively, and then the smoothing proximal gradient descent method [54] is applied to optimize a smooth surrogate of the objective function. A detailed procedure of the optimization algorithm is shown in Algorithm 1.

E. Inference

Once the model parameters $\mathbf{W}$ are learned, we obtain a binary linear classifier for each action class. At inference, given a new observation $\mathbf{X}_n$, we first generate the proper input feature $\widetilde{\mathbf{X}}_n$ in accordance with $\mathcal{J}$, and then calculate decision values across all classes and all joint configurations by running all the class classifier on it. Finally the class that results in the largest decision value is the predicted label $\hat{Y}$

$$\hat{Y} = \underset{(Y,h)\in\mathcal{Y}\times\mathcal{J}}{\arg\max}\left[\mathbf{W}\cdot\psi(\widetilde{\mathbf{X}}_n, h)\right]. \tag{16}$$

IV. EXPERIMENTS

A. Datasets

To evaluate our proposed approach, three benchmark datasets are utilized: UTKinect-Action dataset [12], MSR-Action3D dataset [17] and Florence3D-Action dataset [18]. Table I summarizes the main statistics of three datasets in brief.

UTKinect-Action dataset: In this dataset, there are 10 action classes and 199 action sequences in total. The set of actions consists of *walk*, *sit down*, *stand up*, *pick up*, *carry*, *throw*, *push*, *pull*, *wave hands*, *clap hands*. Each action is performed by 10 different actors twice. The 3D world coordinates of 20 joints are provided for each action sequence. The temporal duration of all sequences varies from 5 to 74 frames. Besides the challenge of this dataset is the high intra-class variations, i.e. some actors throw an object with either their right or left arm.

MSR-Action3D dataset: There are 20 different action classes performed by 10 different actors 2∼3 times, 567 action sequences in total. The set of actions includes *high arm wave*, *horizontal arm wave*, *hammer*, *hand catch*, *forward punch*, *high throw*, *draw x*, *draw tick*, *draw circle*, *hand clap*, *two hand wave*, *side-boxing*, *bend*, *forward kick*, *side kick*, *jogging*, *tennis swing*, *tennis serve*, *golf swing*, *pick up and throw*. The 3D world coordinates of 20 joints are provided with the dataset. The sequences ranges in from 13 to 76 frames which means a large temporal misalignment. Besides the challenge of this dataset is that many of actions are highly similar to each other.

Florence3D-Action dataset: This dataset is captured by a stationary Kinect sensor. 9 actions are comprised in this dataset: *wave, drink*, *answer phone*, *clap*, *tight lace*, *sit down*, *stand up*, *read watch*, *bow*. Each one is performed by 10 different subjects twice or three times, and there are 215 action sequences altogether. The 3D world coordinates of 15 joints are provided with the dataset. The difficulties are high intra-class variations and the presence of the human-object interaction.

TABLE III
COMPARISON WITH THE STATE-OF-THE-ART RESULTS ON UTKINECT-ACTION DATASET

Methods	Recognition Rate
Histograms of 3D joints [12]	90.92
EigenJoints [55]	92.38
Spatial and temporal part-sets [56]	90.22
Cov3DJ [11]	90.53
Random forests [57]	90.90
Points in a Lie Group [24]	97.08
LM^3 TL [31]	98.80
LCSS+MIJA [38]	-
Ker-RP-RBF [22]	-
Proposed method	**99.19**

TABLE IV
COMPARISON WITH THE STATE-OF-THE-ART RESULTS ON FLORENCE3D-ACTION DATASET

Methods	Recognition Rate
Multi-part Bag-of-Poses [18]	82.00
EigenJoints [55]	82.30
Spatial and temporal part-sets [56]	90.22
Cov3DJ [11]	90.53
Random forests [57]	90.90
Points in a Lie Group [24]	90.88
LM^3 TL [31]	93.42
LCSS+MIJA [38]	-
Ker-RP-RBF [22]	-
Proposed method	**95.29**

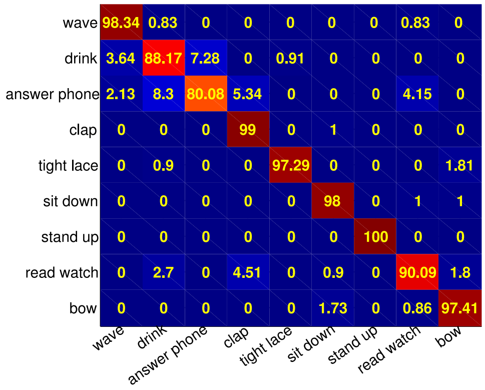


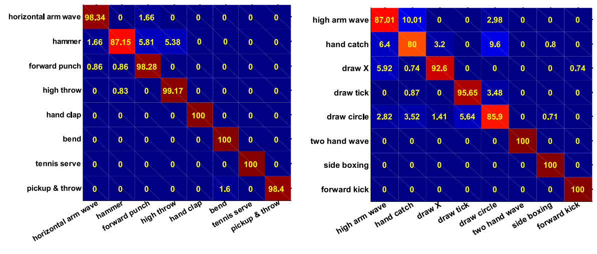
Fig. 2. Confusion matrix for Florence3D-Action dataset.

B. Experimental Settings

In our experiments, all 3D joint coordinates are transformed from 3D world coordinates to a person centric coordinate system by placing the joint *HipCenter* at the origin. To tackle the issue of varying sequence length, dynamic time warping is used to align all the action sequences to the same length. The pairwise relative joint position is used as the feature representation. The original joint configurations set is constructed by combining P joints in all skeleton joints. We evaluate the performance with different P ranging from 2 to 10, and find that $P = 3$ is the best case in terms of both the computational cost as well as the recognition performance. For this case, there are 6 different pairwise combinations for each joint configuration. Therefore, in the stage of MIL, the feature dimension of each instance is $3 \times 6 \times t_f$, where t_f is the time length of the sequence. In the stage of MTL, the final feature representation is concatenated with the feature vectors of these selected joint configurations.

TABLE V
COMPARISON WITH THE STATE-OF-THE-ART RESULTS ON MSR-ACTION3D DATASET

Methods	Recognition Rate
MSR-Action3D dataset (protocal of [17])	
Histograms of 3D joints [12]	78.97
EigenJoints [55]	82.30
Spatial and temporal part-sets [56]	90.22
Cov3DJ [11]	90.53
Random forests [57]	90.90
Points in a Lie Group [24]	92.46
LM^3 TL [31]	95.62
LCSS+MIJA [38]	91.2
Ker-RP-RBF [22]	**96.9**
Proposed approach	96.37
MSR-Action3D dataset (protocal of [13])	
Actionlets [13]	88.20
Points in a Lie Group [24]	89.48
LM^3 TL [31]	90.53
Proposed method	**93.63**



(a) (b)

Fig. 3. Confusion matrices: (a) MSR-Action3D AS_1; (b) MSR-Action3D AS_2.

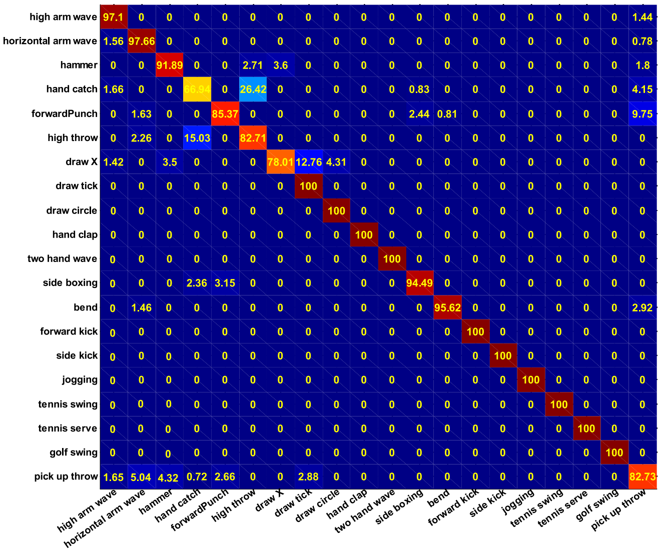
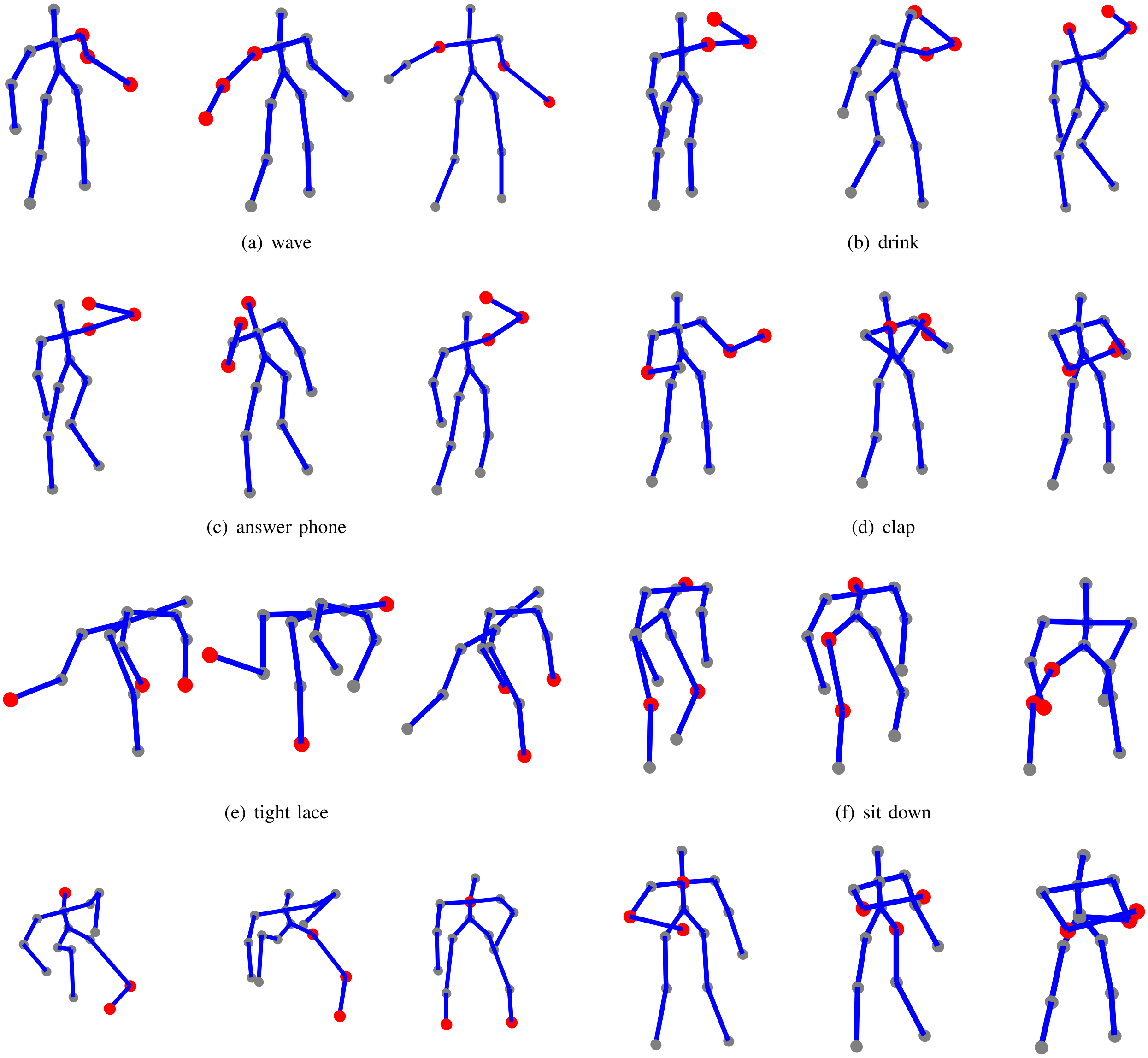


Fig. 4. Confusion matrix for MSR-Action3D dataset.



(a) wave

(b) drink

(c) answer phone

(d) clap

(e) tight lace

(f) sit down

(g) stand up

(h) read watch

Fig. 5. Part visual examples of the learned joint configurations belonging to eight different action class on Florence3D-Action dataset. The joint configurations are marked as red points. (To facilitate visualization, the skeleton sequence frames are rotated.)

We conduct 10-fold cross-validation testing over ten different combinations of training and test data. In each fold, half actors are used to train and the other half to test. The reported results are the average values over these 10-fold cross-validation. For MSR-Action3D dataset, two kinds of validation protocol are following as stated in [17] and [13]. In the former, the dataset are divided into 3 overlapping subsets *AS1*, *AS2*, and *AS3*, each consisting of 8 actions, and performed recognition on each subset separately. And in the latter all 20 actions are involved simultaneously, which is more difficult compared with the former.

In the stage of MIL, the parameter $\mathbf{w}_l$ for l-th action class is initialized with an identity vector. The scalar parameters C, C_d and C_m in (3) are fine tuned by searching the grid of $\{10^{-3}, 10^{-2}, \ldots, 10^{3}\}$ for each action. Therefore those parameters are slightly different for different actions. The parameters λ_1 and λ_2 in (14) follow the same strategy, searching the grid of $\{10^{-5}, 10^{-4}, \ldots, 10^{5}\}$ for the best performance.

We use *accuracy* as the evaluation metric to measure the recognition performance of the action recognition methods, which is the percentage of the correct recognition results in the entire test data.

C. Analysis

1) Effects of Regularization Terms: In our proposed approach, we explore the group structure of the action classes $\{G_k\}_{k=1}^{K}$. And then we integrate this property as a structured sparsity regularization into the MTL model in order to capture the intrinsic correlation between joint configurations and action classes. To evaluate the effectiveness of this sparsity regularization, we compare the average recognition accuracy of our approach with other two baselines, i.e., *Frobenius norm* and ℓ_{21} norm. Here ℓ_{21} *norm* is formulated as $\|\mathbf{W}\|_{2,1} = \sum_{m=1}^{N_s} \|\mathbf{W}^m\|_2$, which encourages to learn individual weight across different joint configurations. For MSR-Action 3D datasets, the two protocols [13], [17] mentioned above are evaluated in MSRA_A and MSRA_S, respectively. As shown in Fig. 1, our proposed method outperforms the other two counterparts. Specifically, our proposed regularization is around 8% superior to the Frobenius norm in the average recognition accuracy on all three datasets. And our average superiority compared with ℓ_{21} norm is 2% for UTKinect dataset, 4% for Florence dataset and MSR-Action 3D datasets, and 3% for MSR-Action 3D datasets. It clearly demonstrates that the mixed norm imposed on our proposed approach is reasonable and effective.

2) Comparison With State-of-the-Arts: In Table II, we summarize the detailed information about the comparison approaches, such as the feature representation and the classifiers, in order to make the comparison more intuitive. In Table III, we report the average recognition accuracy on UTKinect-Action dataset. The average accuracy of the proposed method is 99.19%, our method outperforms the state-of-the-art by 0.39%. The confusion matrix is skipped due to its high classification recognition accuracy.

Table IV gives the results regrading Florence3D-Action dataset. It is shown that the proposed method achieves the best recognition accuracy of 95.29%. The performance of our method is about 1.87% higher than the state-of-the-art approaches. Fig. 2 is the confusion matrix by our method on this dataset to gain insight into the quantitative results.

Table V reports the average recognition accuracy on MSR-Action3D dataset. Specifically, our proposed method achieves the nearly highest recognition accuracy as 96.37% and 93.63% corresponding to the two validation protocols, as stated in [13] and [17], respectively. It outperforms the state-of-the-art approaches by 3.1% on average. Fig. 3 shows the confusion matrices for MSR-Action3D AS_1 and MSR-Action3D AS_2 datasets. We can see that most of the confusion are between highly similar actions such as *high throw* in MSR-Action3D AS_1 dataset, *draw X*, *hand catch*, and *side boxing* in the case of MSR-Action3D AS_2. Fig. 4 illustrates the confusion matrix for MSR-Action3D dataset. We can see that the proposed method performs very well on most of the actions, except some actions like *hand catch* and *high throw*, since these actions are too similar to distinguish them effectively.

3) Visual Examples of Learned Joint Configuration: Fig. 5 shows partial examples of the mined joint configurations for different action classes on Florence3D-Action dataset. From the visualization, we can see that these joint configurations for the same action class have the property of diversity, which can be used to cope with the intra-class variations. For example, *wave* is various from person to person, i.e., some actors wave with either their right or left arm, while others wave with both arms. Meanwhile, for the different action classes, such as *drink* and *answer phone*, the discriminative joint configurations can efficiently solve the problem the inter-class ambiguities.

V. Conclusion

We presented a two-stage multi-instance multitask learning framework, termed MIMTL, for 3D human action recognition by distinguishing a certain action class with a set of discriminative joint configurations. Specifically, multi-instance learning model is first utilized to discover the discriminative joint configurations automatically, in which the action class and the corresponding joint configurations are regarded as the bag and its instances. Then, multitask learning model is applied to relate the joint configurations across different action classes. Due to the fact that the action classes have the property of group structure, we integrate a mixed group sparsity constraint into the multitask learning model to enhance the discriminant abilities of the joint configurations for a certain action class. Extensive experiments on three benchmark datasets demonstrate encouraging performance over a number of state-of-the-art approaches.

In the future, we plan to extend our proposed approach to large-scale 3D action recognition datasets, such as [58]. In addition, deep learning is appealing in many different applications, which is more applicable to large-scale dataset. For action recognition community, deep learning is mainly exploited to learn visual features by using a deep neural network trained from a large number of labeled data, and typical methods include 3D ConvNets [59], Two-Stream ConvNets [60], and trajectory-pooled deep-convolutional descriptor [61]. For skeleton based action recognition, [62] divides the human skeleton into five parts, and then separately feeds them to five bidirectional recurrent neural networks (BRNNs). Inspired by them, we will consider to design a deep multitask learning model to discriminatively capture the intrinsic relationship between joint configurations and action class.

References

[1] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld, “Learning realistic human action from movies,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2008, pp. 1–8.

[2] L. Wang, Y. Qiao, and X. Tang, “Mining motion atoms and phrases for complex action recognition,” in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 2680–2687.

[3] D. Wu and L. Shao, “Leveraging hierarchical parametric networks for skeletal joints based action segmentation and recognition,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2014, pp. 724–731.

[4] J. Wang and Y. Wu, “Learning maximum margin temporal warping for action recognition,” in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 2688–2695.

[5] J. Wang, X. Nie, Y. Xia, Y. Wu, and S.-C. Zhu, “Cross-view action modeling, learning and recognition,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2014, pp. 2649–2656.

[6] F. Ofli, R. Chaudhry, G. Kurillo, R. Vidal, and R. Bajcsy, “Sequence of the most informative joints (SMIJ): A new representation for human skeletal action recognition,” *J. Vis. Commun. Image Represent.*, vol. 25, no. 1, pp. 24–38, 2014.

[7] Y. Yan, E. Ricci, R. Subramanian, G. Liu, and N. Sebe, “Multitask linear discriminant analysis for view invariant action recognition,” *IEEE Trans. Image Process.*, vol. 23, no. 12, pp. 5599–5611, Dec. 2014.

[8] Y. Zhou, B. Ni, R. Hong, M. Wang, and Q. Tian, “Interaction part mining: A mid-level approach for fine-grained action recognition,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2015, pp. 3323–3331.

[9] J. Luo, W. Wang, and H. Qi, “Group sparsity and geometry constrained dictionary learning for action recognition from depth maps,” in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 1809–1816.

[10] J. Shotton *et al.*, “Real-time human pose recognition in parts from single depth cameras,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2011, pp. 1297–1304.

[11] M. Hussein, M. Torki, M. Gowayyed, and M. El-Saban, “Human action recognition using a temporal hierarchy of covariance descriptors on 3D joint locations,” in *Proc. 23rd Int. Joint Conf. Artif. Intell.*, 2013, pp. 2466–2472.

[12] L. Xia, C. Chen, and J. Aggarwal, “View invariant human action recognition using histograms of 3D joints,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recog. Workshops*, Jun. 2012, pp. 20–27.

[13] J. Wang, Z. Liu, Y. Wu, and J. Yuan, “Mining actionlet ensemble for action recognition with depth cameras,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2012, pp. 1290–1297.

[14] Y. Yang, Z. Ma, Z. Xu, S. Yan, and A. Hauptmann, “How related exemplars help complex event detection in web videos,” in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 2104–2111.

[15] Z. Ma, Y. Yang, N. Sebe, and A. Hauptmann, “Knowledge adaptation with partially shared features for event detection using few exemplars,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 9, pp. 1789–1802, Sep. 2014.

[16] Y. Yang, Z. Ma, F. Nie, X. Chang, and G. Alexander, “Multi-class active learning by uncertainty sampling with diversity maximization,” *Int. J. Comput. Vis.*, no. 113, pp. 113–127, 2015.

[17] W. Li, Z. Zhang, and Z. Liu, “Action recognition based on a bag of 3D points,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recog. Workshops*, Jun. 2010, pp. 9–14.

[18] L. Seidenari, V. Varano, S. Berretti, A. D. Bimbo, and P. Pala, “Recognizing actions from depth cameras as weakly aligned multi-part bag-of-poses,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recog. Workshops*, Jun. 2013, pp. 479–485.

[19] Y. Sheikh, M. Sheikh, and M. Shah, “Exploring the space of a human action,” in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2005, vol. 1, pp. 144–149.

[20] F. Lv and R. Nevatia, “Recognition and segmentation of 3D human action using HMM and multi-class AdaBoost,” in *Proc. 9th Eur. Conf. Comput. Vis.*, 2006, pp. 359–372.

[21] R. Chaudhry, F. Ofli, G. Kurillo, R. Bajcsy, and R. Vidal, “Bio-inspired dynamic 3D discriminative skeletal features for human action recognition,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recog. Workshops*, Jun. 2013, pp. 471–478.

[22] L. Wang, J. Zhang, L. Zhou, C. Tang, and W. Li, “Beyond covariance: Feature representation with nonlinear kernel matrices,” in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2015, pp. 4570–4578.

[23] X. Yang and Y. Tian, “Eigenjoints-based action recognition using naive-Bayes-nearest-neighbour,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2012, pp. 14–19.

[24] R. Vemulapalli, F. Arrate, and R. Chellappa, “Human action recognition by representing 3D skeletons as points in a lie group,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2014, pp. 588–595.

[25] M. Devanne *et al.*, “3D human action recognition by shape analysis of motion trajectories on Riemannian manifold,” *IEEE Trans. Cybern.*, vol. 45, no. 7, pp. 1340–1352, Jul. 2015.

[26] A. Sharaf, M. Torki, M. Hussein, and M. El-Saban, “Real-time multi-scale action detection from 3D skeleton data,” in *Proc. CACV Conf.*, 2015, pp. 998–1005.

[27] M. Zanfir, M. Leordeanu, and C. Sminchisescu, “The moving pose: An efficient 3D kinematics descriptor for low-latency action recognition and detection,” in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 2752–2759.

[28] H. Pazhoumand-Dar, C. Lam, and M. Masek, “Joint movement similarities for robust 3d action recognition using skeletal data,” *J. Vis. Commun. Image Represent.*, no. 30, pp. 10–21, 2015.

[29] P. Wei, N. Zheng, Y. Zhao, and S. Zhu, “Concurrent action detection with structural prediction,” in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 3136–3143.

[30] A. Eweiwi, M. Cheema, C. Baukhage, and J. Gall, “Efficient pose-based action recognition,” in *Proc. 12th Asian Conf. Comput. Vis.*, 2014, pp. 428–443.

[31] Y. Yang *et al.*, “Latent max-margin multitask learning with skelets for 3D action recognition,” *IEEE Trans. Cybern.*, to be published.

[32] J. Frank, “A review of multi-instance learning assumptions,” *Knowl. Eng. Rev.*, vol. 25, pp. 1–25, 2010.

[33] T. Dietterich, R. Lathrop, and T. Lozano-Perez, “Solving the multiple instance problem with axis-parallel rectangles,” *Artif. Intell.*, vol. 89, nos. 1/2, pp. 31–71, 1997.

[34] Q. Zhang, S. A. Goldman, W. Yu, and J. Fritts, “Content-based image retrieval using multiple-instance learning,” in *Proc. 19th Int. Conf. Mach. Learn.*, 2002, pp. 682–689.

[35] O. Maron and A. Ratan, “Multiple-instance learning for natural scene classification,” in *Proc. 15th Int. Conf. Mach. Learn.*, 1998, pp. 341–349.

[36] K. Lai, F. Yu, M. Chen, and S. Chang, “Video event detection by inferring temporal instance labels,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2014, pp. 2251–2258.

[37] W. Li, L. Duan, D. Xu, and I. Tsang, “Text-based image retrieval using progressive multi-instance learning,” in *Proc. IEEE Int. Conf. Comput. Vis.*, Nov. 2011, pp. 2049–2055.

[38] H. Hajimirsadeghi, Y. Wang, A. Vahdat, and G. Mori, “Visual recognition by counting instances: A multi-instance cardinality potential kernel,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2015, pp. 2596–2605.

[39] A. Argyriou, T. Evgeniou, and M. Pontil, “Convex multitask feature learning,” *Mach. Learn.*, vol. 73, no. 3, pp. 243–272, 2008.

[40] J. Liu, S. Ji, and J. Ye, “multitask feature learning via efficient L2,1-norm minimization,” in *Proc. 25th Conf. Uncertainty Artif. Intell.*, 2009, pp. 339–348.

[41] S. Parameswaran and K. Q. Weinberger, “Large margin multitask metric learning,” in *Proc. 23rd Int. Conf. Neural Inf. Process. Syst.*, 2010, pp. 1867–1875.

[42] S. Kim and E. P. Xing, “Tree-guided group lasso for multitask regression with structured sparsity,” in *Proc. 27th Int. Conf. Mach. Learn.*, 2010, pp. 543–550.

[43] L. Jacob, F. Bach, and J. P. Vert, “Clustered multitask learning: A convex formulation,” in *Proc. Adv. Neural Inf. Process. Syst. Conf.*, 2009, vol. 21, pp. 745–752.

[44] J. Zhou, J. Chen, and J. Ye, “Clustered multitask learning via alternating structure optimization,” in *Proc. 24th Int. Conf. Neural Inf. Process. Syst.*, 2011, pp. 702–710.

[45] A. Kumar and H. Iii, “Learning task grouping and overlap in multitask learning,” in *Proc. 29th Int. Conf. Mach. Learn.*, 2012, pp. 1383–1390.

[46] D. Jayaraman, F. Sha, and K. Grauman, “Decorrelating semantic visual attributes by resisting the urge to share,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2014, pp. 1629–1636.

[47] B. Mahasseni and S. Todorovic, “Latent multitask learning for view-invariant action recognition,” in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 3128–3135.

[48] Q. Zhou, G. Wang, K. Jia, and Q. Zhao, “Learning to share latent tasks for action recognition,” in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 2264–2271.

[49] Y. Lin, J. Hua, N. Tang, M. Chen, and H. M. Liao, “Depth and skeleton associated action recognition without online accessible RGB-D cameras,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2014, pp. 2617–2624.

[50] R. Hou, A. Zamir, R. Sukthankar, and M. Shah, “DaMN—Discriminative and mutually nearest: Exploiting pairwise category proximity for video action recognition,” in *Proc. 13th Eur. Conf. Comput. Vis.*, 2014, pp. 721–736.

[51] Y. Yang, R. Liu, C. Deng, and X. Gao, “multitask human action recognition via exploring super-category,” *Signal Process.*, vol. 124, pp. 36–44, 2016.

[52] C.-N. J. Yu and T. Joachims, “Learning structural SVMs with latent variables,” in *Proc. 26th Annu. Int. Conf. Mach. Learn.*, 2009, pp. 1169–1176.

[53] S. Kim and E. P. Xing, “Tree-guided group lasso for multi-response regression with structured sparsity, with an application to eQTL mapping,” *Ann. Appl. Statist.*, vol. 6, no. 3, pp. 1095–1117, 2012.

[54] X. Chen, Q. Lin, S. Kim, J. G. Carbonell, and E. P. Xing, “Smoothing proximal gradient method for general structured sparse learning,” *CoRR*, 2012. [Online]. Available: http://arxiv.org/abs/1202.3708

[55] X. Yang and Y. Tian, “Effective 3D action recognition using Eigen Joints,” *J. Vis. Commun. Image Represent.*, vol. 25, no. 1, pp. 2–11, 2014.

[56] C. Wang, Y. Wang, and A. L. Yuille, “An approach to pose based action recognition,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2013, pp. 915–922.

[57] Y. Zhu, W. Chen, and G. Guo, “Fusing spatiotemporal features and joints for 3D action recognition,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recog. Workshops*, Jun. 2013, pp. 486–491.

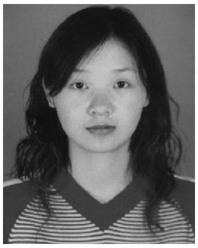
[58] A. Shahroudy, J. Liu, T.-T. Ng, and G. Wang, “NTU RGB+D: A large scale dataset for 3D human activity analysis,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2016, pp. 1010–1019.

[59] S. Ji, W. Xu, M. Yang, and K. Yu, "3D convolutional neural networks for human action recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 1, pp. 221–231, Jan. 2013.

[60] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," in *Proc. Adv. Neural Inf. Process. Syst. Conf.*, 2014, pp. 568–576.

[61] L. Wang, Y. Qiao, and X. Tang, "Action recognition with trajectory-pooled deep-convolutional descriptors," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2015, pp. 4305–4314.

[62] Y. Du, W. Wang, and L. Wang, "Hierarchical recurrent neural network for skeleton based action recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2015, pp. 1110–1118.



Yanhua Yang (M'14) received the B.E. degree in electronic and information engineering and the M.S. degree in signal and information processing in 2004 and 2007, respectively, from Xidian University, Xi'an, China, where she is currently working toward the Ph.D. degree in electronic engineering.

Her main research interests include complex action recognition and event detection.

Cheng Deng (M'10) received the B.E., M.S., and Ph.D. degrees in signal and information processing from Xidian University, Xi'an, China, in 2001, 2006, and 2009, respectively.

He is currently a Full Professor with the School of Electronic Engineering, Xidian University. He has authored or coauthored more than 50 scientific articles at top venues, including the IEEE Transactions on Neural Networks and Learning Systems, the IEEE Transactions on Multimedia, the IEEE Transactions on Cybernetics, the IEEE Transactions on Systems, Man, and Cybernetics, the IEEE Transactions on Image Processing, the IEEE International Conference on Computer Vision, the IEEE Conference on Computer Vision and Pattern Recognition, International Joint Conference on Artificial Intelligence, and AAAI Conferences. His research interests include computer vision, multimedia processing and analysis, and information hiding.

Shangqian Gao received the B.E. degree in electronic and information engineering from Xidian University, Xi'an, China, in 2004, and is currently working toward the M.S. degree at the College of Engineering, Northeastern University, Boston, MA, USA.

His main research interests include computer vision and machine learning.

Wei Liu is currently a Full Professor with the School of Electronic Engineering, Xidian University, Xi'an, China. His research interests include computer vision, image processing, machine learning, and pattern recognition.

Dapeng Tao (M'14) received the B.E. degree from Northwestern Polytechnical University, Xi'an, China, in 1999, and the Ph.D. degree from the South China University of Technology, Guangzhou, China, in 2014.

He is currently a Full Professor with the School of Information Science and Engineering, Yunnan University, Kunming, China. He has authored or coauthored more than 30 scientific articles. Over the past years, his research interests have included machine learning, computer vision, and robotics.

Prof. Tao has served more than ten international journals, including the IEEE Transactions on Neural Networks and Learning Systems, the IEEE Transactions on Multimedia, the IEEE Transactions on Circuits and Systems for Video Technology, the IEEE Signal Processing Letters, and *Information Sciences*.

Xinbo Gao (M'02–SM'07) received the B.E., M.S., and Ph.D. degrees in signal and information processing from Xidian University, Xi'an, China, in 1994, 1997, and 1999, respectively.

From 1997 to 1998, he was a Research Fellow with the Department of Computer Science, Shizuoka University, Shizuoka, Japan. From 2000 to 2001, he was a Postdoctoral Research Fellow with the Department of Information Engineering, the Chinese University of Hong Kong, Hong Kong, China. Since 2001, he has been with the School of Electronic Engineering, Xidian University. He is currently a Cheung Kong Professor of Ministry of Education, a Professor of Pattern Recognition and Intelligent System, and the Director of the State Key Laboratory of Integrated Services Networks, Xi'an, China. He has authored or coauthored five books and around 200 technical articles in refereed journals and proceedings. His current research interests include multimedia analysis, computer vision, pattern recognition, machine learning, and wireless communications.

Prof. Gao is a Fellow of the Institution of Engineering and Technology. He is on the Editorial Boards of several journals, including *Signal Processing* (Elsevier) and *Neurocomputing* (Elsevier). He served as the General Chair/Co-Chair, Program Committee Chair/Co-Chair, or PC Member for around 30 major international conferences.